

ABSTRACT OF THE DISCLOSURE

A variable latency cache memory is disclosed. The cache memory includes a plurality of storage elements for storing stack memory data in a first-in-first-out manner. The cache memory distinguishes between pop and load instruction requests and provides pop data faster than load data by speculating that pop data will be in the top cache line of the cache. The cache memory also speculates that stack data requested by load instructions will be in the top one or more cache lines of the cache memory. Consequently, if the source virtual address of a load instruction hits in the top of the cache memory, the data is speculatively provided faster than the case where the data is in a lower cache line or where a full physical address compare is required or where the data must be provided from a non-stack cache memory in the microprocessor, but slower than pop data.